Changing social norm compliance with noninvasive brain stimulation

Ruff, Christian C ; Ugazio, Giuseppe ; Fehr, Ernst

Abstract: All known human societies have maintained social order by enforcing compliance with social norms. The biological mechanisms underlying norm compliance are, however, hardly understood. We show that the right lateral prefrontal cortex (rLPFC) is involved in both voluntary and sanction-induced norm compliance. Both types of compliance could be changed by varying the neural excitability of this brain region with transcranial direct current stimulation, but they were affected in opposite ways, suggesting that the stimulated region plays a fundamentally different role in voluntary and sanction-based compliance. Brain stimulation had a particularly strong effect on compliance in the context of socially constituted sanctions, whereas it left beliefs about what the norm prescribes and about subjectively expected sanctions unaffected. Our findings suggest that rLPFC activity is a key biological prerequisite for an evolutionarily and socially important aspect of human behavior.

DOI: https://doi.org/10.1126/science.1241399

Title: Changing Social Norm Compliance With Noninvasive Brain Stimulation

Authors: C.C. Ruff [1]*, G.Ugazio[1,2], E.Fehr[1]*

Affiliations:

[1]Laboratory for Social and Neural Systems Research (SNS-Lab), Department of Economics, University of Zurich, Zurich, Switzerland.

[2]Social, Cognitive and Affective Neuroscience Unit, University of Vienna, Vienna, Austria.

*Correspondence to: christian.ruff@econ.uzh.ch; ernst.fehr@econ.uzh.ch.

All known human societies have maintained social order by enforcing compliance with social norms. The biological mechanisms underlying norm compliance are, however, hardly understood. We show that the right lateral prefrontal cortex is involved in both voluntary and sanction-induced norm compliance. Both types of compliance could be changed by varying neural excitability of this brain region with transcranial direct current stimulation, but they were affected in opposite ways, suggesting that the stimulated region plays a fundamentally different role in voluntary and sanction-based compliance. Brain stimulation had a particularly strong effect for compliance based on socially-constituted sanctions, while it left beliefs about what the norm prescribes and about subjectively expected sanctions unaffected. Our findings suggest that rLPFC activity is a key biological prerequisite for an evolutionarily and socially important aspect of human behavior.

One Sentence Summary:

Human compliance with social norms can be increased or decreased by appropriately stimulating the right lateral prefrontal cortex non-invasively with electrical currents.

Human societies depend crucially on social norms that specify the range of permissible actions for a given situation. Social norms range from the mundane (e.g., dress codes, table etiquette) to the profound (e.g., collective action, bilateral exchange, law obedience). They are considered a hallmark of human civilization because no other known species regulates social interactions to the same degrees by norms (*1-3*). The potential of norms to guide collective behavior can break down if norm violations are not sanctioned, because humans tend to follow prevailing norms conditional on observing others' compliance (*4*). All known human societies have therefore enforced norm compliance by threatening norm violators with punishment, both officially via legal codes and institutions, and informally in the context of private sanctions through peers (*5, 6*). The importance of credible sanctioning threats for maintaining norm compliance is well established by ethnographic evidence (*1, 2*), evolutionary theory (*1, 3*), and laboratory experiments (*5, 6*).

It has been proposed that the human brain may have developed neural processes that support norm enforcement by generating appropriate behavioral responses to social punishment threats (*7-10*). However, neuroscience studies on social norms have mostly focused on the neural basis of punishing others (*11-14*), whereas evidence for neural circuitry underlying sanction-induced compliance with norms is scarce. In mature adults, a brain network involving an area in the right lateral prefrontal cortex (rLPFC) is activated during norm-compliant behavior triggered by social punishment threats (*10*). However, it is not possible to conclude from correlative fMRI findings that norm compliance depends *causally* on neural activity in the rLPFC (*15*). Establishing such a causal dependence is crucial for our understanding of how social norm compliance develops in the context of brain maturation (*16*) and how it is pathologically altered and therapeutically amenable in the context of brain disorders (*9*).

We employed transcranial direct current stimulation (tDCS) (*17*) to examine whether social norm compliance depends *causally* on neural processing in the previously-identified rLPFC region (*10*). Participants engaged via computer terminals in anonymous social interactions that had real financial consequences. In every round, participants ("Player A") received an amount of Money Units (MUs) and decided how much of it to transfer to a randomly assigned anonymous opponent ("Player B"). In *baseline* rounds, this transfer was implemented, whereas in *punishment* rounds, Player B could respond to the transfer by reducing Player A's MUs (Fig. 1, Fig. S1, Supporting Online Material, SOM, (*18*)). In Western Cultures, a fairness norm (*19-21*) prescribes to split the "cake" of MUs equally between both players. This conflicts with Player A's self-interest motive to keep as many MUs as possible. In baseline rounds, Player A thus typically transfers only around 10% - 25% of the MUs. In contrast, when a sanctioning threat is present, Player A largely obeys the fairness norm and transfers around 40% - 50% of the MUs (*10, 20*). The transfer difference between punishment and baseline rounds thus indexes *sanction-induced* norm compliance, i.e., the degree to which the sanction threat induces Player A to change her transfer from the level of *voluntary* norm-compliance as measured in baseline rounds.

Individual differences in sanction-induced norm compliance correlate with fMRI-measured activity in the rLPFC (*10*). Based on this finding and the rLPFC's general role in the control of behavior (*22, 23*), it has been proposed that the rLPFC may weigh fair versus selfish responses specifically when punishment threats are present (*8, 10*). To provide causal evidence for this hypothesis, we first identified the specific rLPFC region described in (*10*) using MR-scans of 63 female participants; we then experimentally altered neural excitability in this brain area during behavioral performance in a double-blind, placebo-controlled tDCS design (SOM,

Fig. S2). tDCS can both increase or decrease neural excitability in the stimulated region, depending on the polarity of the current flow (*17*). We thus randomly sorted participants into three stimulation groups where neural excitability in the rLPFC was enhanced with anodal tDCS, reduced with cathodal tDCS, or left unaltered by sham/placebo tDCS as control for possible non-neural effects of stimulation (see SOM). Such non-neural effects did not differ between the groups (see SOM) and therefore could not account for performance in the norm-compliance paradigm.

Participants were sensitive to the punishment threat and transferred more money in punishment than in baseline rounds (mean transfer difference 29.44 MUs; $p < 0.001$, GLS regression). However, in line with our hypothesis, the two active brain stimulation conditions changed sanction-induced norm compliance in opposite ways relative to the sham condition (Fig. 2A, Table S2). Anodal tDCS *increased* the transfer difference by 33.5% (GLS regression, $p < 0.001$) whereas cathodal tDCS *decreased* the transfer difference by 22.7% ($p < 0.001$).

Do these effects reflect changes in altruistic behavior, with increased (decreased) monetary transfers regardless of punishment threats? This interpretation is refuted by the data on *voluntary norm-compliance* in baseline rounds (Fig. 2B, Table S3). Voluntary transfers were actually *decreased* (GLS regression, $p < 0.001$) during anodal tDCS and *increased* ($p < 0.01$) during cathodal tDCS, relative to the sham condition. This not only confirms that tDCS affected subjects' response to the punishment threat but that these tDCS effects on sanction-induced compliance were actually stronger than the opposite effects on voluntary compliance: If tDCS had not affected sanction-induced compliance then overall transfers in punishment rounds – which are based on voluntary plus sanction-induced compliance – should also be lower after anodal and higher after cathodal stimulation. However, overall transfers in punishment rounds

were in fact *higher* (GLS regression, $p < 0.05$) during anodal tDCS and *lower* ($p < 0.001$) during cathodal tDCS than in the sham condition (Fig. S3).

Which task-related psychological mechanisms may have contributed to the tDCS effect? To respond appropriately, participants need to know the fairness norm and form appropriate beliefs about Player B's reactions. We measured (i) the participants' perceived fairness, (ii) the anger they expected the opponent to feel, and (iii) the punishment they expected at different transfer levels (Fig. 3). All participants were clearly aware of the fairness norm and rated higher transfers as significantly fairer (ANOVA, $F(2,60) = 84.88$, $p < 0.001$), less likely to cause anger in the opponent ($F(2,60)=218.96$, $p < 0.001$), and leading to lower punishment ($F(2,60) = 82.69$, $p < 0.001$). Importantly, the type of brain stimulation did *not* affect participants' beliefs, neither on average (all $F(2,60) < 0.94$, all $p > 0.39$) nor in their change across different transfer levels (all $F(2,60) < 0.55$, all $p > 0.74$).

Our findings do not yet show that the stimulated rLPFC region implements specifically social aspects of behavioral control. In particular, behavior in punishment rounds requires risk taking and trading off higher transfers with a lower risk of sanction. We therefore repeated the experiment in a sample of 59 new female volunteers who took the identical decisions as before, but now played against a computer pre-programmed to respond in the same way as a human opponent in punishment rounds (see SOM). In this "non-social context", participants were also sensitive to punishment threats (Fig. S4A) but the effects of tDCS on sanction-induced transfers were significantly weaker than during interactions with human opponents (Fig. 4A and Table S3). This held for both increases in sanction-induced transfers due to anodal tDCS (GLS regression, $p = 0.009$) and decreases due to cathodal tDCS ($p = 0.001$, GLS regression). In baseline rounds of the non-social context – where no social norm prescribes sharing MUs with

the computer – participants hardly transferred any MUs (Fig. S4B). Such (possibly erroneous) *voluntary* transfers to the computer were therefore also less affected by tDCS than norm-related voluntary transfers to human opponents (Fig. 4B; GLS regression, $p < 0.05$ for anodal tDCS and $p < 0.001$ for cathodal tDCS).

Social punishment is thought to have played an important role for the evolution of human social behavior and cooperation (*1-3*). Our results show that the influence of punishment threats on human social norm compliance depends causally on neural activity in the rLPFC. This suggests a neural mechanism involving the rLPFC that aligns behavior with social norms when punishment is possible. The more pronounced involvement of this mechanism for genuinely social punishments concurs with suggestions that during human brain evolution, the steep increase in the complexity of social interactions may have shaped specific neural processes for social behavior (*8, 24*). That tDCS affected sanction-induced and voluntary norm compliance in opposite ways suggests that these two forms of norm compliance involve distinct neural circuits; in particular, the rLPFC seems to play a fundamentally different role in voluntary and sanction-based norm compliance.

Our finding that rLPFC stimulation did not affect awareness of the fairness norm and expected sanctions suggests that the rLPFC process necessary for norm-compliant behavior is dissociated from neural mechanisms enabling humans to anticipate sanctions for norm violations and to distinguish “right” from “wrong”. The rLPFC mechanism necessary for norm-compliance is probably not restricted to neural activity within this brain area, given that prefrontal cortex is involved in many aspects of behavioral control (*23*) and that brain stimulation can affect areas interconnected with the stimulation site (*25*). The anatomical connectivity (*26*) and context-dependent functions of prefrontal cortex (*27*) make it more likely that the stimulated rLPFC area

integrates and coordinates activity in a network of brain regions triggered by the need for considering social punishments during action control (*8*).

Brain stimulation studies in humans have so far mostly shown unidirectional, maladaptive effects on decision making, rendering participants more impulsive (*28*), selfish (*29*), or cognitively biased (*30*). Such interventions may therefore be of limited practical use in applied settings. Our finding that changes in the neural excitability of rLPFC can enhance voluntary and sanction-induced social norm compliance may be of relevance because non-compliance with social norms constitutes a major problem in psychiatric (41) and neurological (*31, 32*) disorders, during abnormal development in adolescence (*33*), and in adults in the form of criminal activity (*9*). However, the opposite influence of brain stimulation on voluntary and sanction-induced norm compliance also suggests that increasing one type of norm compliance with brain stimulation may come at the cost of decreasing the other type.

References and Notes:

1. E. Sober, D. S. Wilson, *Unto Others - The Evolution and Psychology of Unselfish Behavior*. (Harvard University Press, Cambridge, Massachusetts, 1998), pp. 394.
2. S. Mathew, R. Boyd, *Proceedings of the National Academy of Sciences of the United States of America* 108, 11375 (Jul 12, 2011).
3. R. Boyd, H. Gintis, S. Bowles, P. J. Richerson, *P Natl Acad Sci USA* 100, 3531 (MAR 18, 2003).
4. U. Fischbacher, S. Gächter, E. Fehr, *Econ Lett* 71, 397 (2001).
5. R. Forsythe, H. L. Horowitz, N. E. Savin, M. Sefton, *Game Econ Behav* 6, 347 (1994).
6. E. Fehr, S. Gächter, *Nature* 415, 137 (2002).
7. P. R. Montague, T. Lohrenz, *Neuron* 56, 14 (Oct 4, 2007).
8. J. W. Buckholtz, R. Marois, *Nat Neurosci* 15, 655 (May, 2012).
9. A. Raine, Y. Yang, *Soc. Cogn. Affect. Neurosci* 1, 203 (2006).
10. M. Spitzer, U. Fischbacher, B. Herrnberger, G. Gron, E. Fehr, *Neuron* 56, 185 (Oct 4, 2007).
11. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *Science* 300, 1755 (2003).
12. D. J. de Quervain *et al.*, *Science* 305, 1254 (2004).
13. J. W. Buckholtz *et al.*, *Neuron* 60, 930 (Dec 10, 2008).
14. M. J. Crockett *et al.*, *J Neurosci* 33, 3505 (Feb 20, 2013).
15. For example, individuals with particular personality traits - such as anxious individuals or those with the tendency for Machiavellian behaviour - may have generally higher LPFC activity and a generally higher propensity to respond to sanctions, but apart from their co-variation, these two variables may not directly influence one another. Alternatively, it is also possible that rather than being the cause of norm compliance, brain activation in rLPFC is merely the consequence of norm compliance. In other words, individuals who respond more strongly to the sanctioning threat may recruit rLPFC more strongly as a result of their choice.
16. N. Steinbeis, B. C. Bernhardt, T. Singer, *Neuron* 73, 1040 (Mar 8, 2012).
17. M. A. Nitsche, W. Paulus, *J Physiol-London* 527, 633 (Sep 15, 2000).
18. Methods and materials, supplemental analyses and supplementary figures are available as supporting material on Science Online.
19. E. Fehr, U. Fischbacher, *Evolution and Human Behavior* 25, 63 (MAR, 2004).
20. C. F. Camerer, *Behavioral game theory: Experiments in strategic interaction*. (Princeton University Press, Princeton, N.J., 2003), pp. 550.
21. J. Henrich *et al.*, *Am Econ Rev* 91, 73 (2001).
22. A. R. Aron, T. W. Robbins, R. A. Poldrack, *Trends Cogn Sci* 8, 170 (Apr, 2004).
23. E. K. Miller, J. D. Cohen, *Annu.Rev.Neurosci.* 24, 167 (2001).
24. R. I. M. Dunbar, *Trends Cogn Sci* 16, 101 (Feb, 2012).
25. J. Driver, F. Blankenburg, S. Bestmann, W. Vanduffel, C. C. Ruff, *Trends Cogn Sci* 13, 319 (Jul, 2009).
26. P. L. Croxson *et al.*, *J Neurosci* 25, 8854 (Sep 28, 2005).
27. J. Duncan, *Trends Cogn Sci* 14, 172 (Apr, 2010).
28. B. Figner *et al.*, *Nat Neurosci* 13, 538 (May, 2010).

29. D. Knoch, A. Pascual-Leone, K. Meyer, V. Treyer, E. Fehr, *Science* 314, 829 (Nov 3, 2006).
30. G. Xue, C. H. Juan, C. F. Chang, Z. L. Lu, Q. Dong, *P Natl Acad Sci USA* 109, 4401 (Mar 20, 2012).
31. A. Bechara, M. Van Der Linden, *Curr Opin Neurol* 18, 734 (Dec, 2005).
32. B. King-Casas *et al.*, *Science* 321, 806 (Aug 8, 2008).
33. D. P. Farrington, R. Loeber, *Child Adol Psych Cl* 9, 733 (Oct, 2000).
34. M. B. Iyer *et al.*, *Neurology* 64, 872 (Mar 8, 2005).
35. B. Fritsch *et al.*, *Neuron* 66, 198 (Apr 29, 2010).

Acknowledgments: C.C.R. and E.F. designed research; C.C.R. and G.U. conducted research; C.C.R. and G.U. analyzed data with input from E.F.; C.C.R. and E.F. wrote the paper with comments by G.U. We thank Karl Treiber and Anthony Schlaepfer for their assistance with data collection. This study was supported by an ERC grant to EF (Foundations of Economic Preferences), Swiss National Science Foundation (SNSF) grants to CCR and EF (CRSII3_141965 and 51NF40_144609), and the SNSF National Competence Center for Research (NCCR) in Affective Sciences.

Figure Captions:

Fig. 1. Economic game used to measure social norm compliance. In each round, both players receive 25 money units (MUs). Player A is given an additional 100 MUs that she can share with Player B by sending a transfer X (in multiples of 10 MUs). All experimental MUs are exchanged into real money at the end of the experiment. Two types of rounds are presented in random order. **(A)** Baseline round: Transfer X is implemented as proposed, measuring Player A's *voluntary* norm compliance. **(B)** Punishment round: Player B can either accept X (blue font) or invest Y MUs from her initial endowment to punish Player A (red font). Y can be any integer between 0 and 25, reducing A's payoff by 5*Y MUs. Player A is aware of this possible sanction; any increase in transfers for punishment relative to baseline rounds therefore measures *sanction-induced* norm compliance.

Fig. 2. rLPFC stimulation changes sanction-induced and voluntary norm compliance. **(A)** Sanction-induced norm-compliance: Average (+/- s.e.m.) transfer difference for punishment rounds minus baseline rounds. Higher values indicate that the punishment threat led to a larger adjustment of transfers towards the fairness norm of an equal split. **(B)** Voluntary norm compliance: Average (+/- s.e.m.) transfers for baseline rounds. All values determined with regression in eq.1 (SOM) ; * $p < 0.05$.

Fig. 3 rLPFC stimulation does not affect participants' beliefs about the fairness of different transfers and about Player B's anticipated anger and expected punishment. **(A)** Average rating of perceived fairness for different transfer levels (scale from 1/"very unfair" to 4/"very fair"). **(B)**

Average rating of anticipated anger felt by Player B for different transfer levels (scale from 1/"not angry at all" to 4/"very angry"). **(C)** Average expected payoff reduction resulting from B's punishment. Error bars represent s.e.m.

Fig. 4. rLPFC stimulation effects are stronger during social interactions. **(A)** tDCS effects on sanction-induced norm compliance during interactions with a human (Social Context) or a computer opponent (Non-social Context). Bars depict average changes in transfer difference for anodal and cathodal tDCS relative to the sham condition. **(B)** tDCS-related changes of voluntary transfers in baseline rounds. Bars represent average changes for anodal and cathodal tDCS relative to the sham condition**.** All values determined with regression in eq. 2 (SOM); * $p < 0.05$.

Supplementary Materials:

Materials and Methods

- Participants and Procedure
- Experimental Paradigm and Measures
- Transcranial Direct Current Stimulation (tDCS)
- Analysis and Results

Figures S1-S3

Tables S1-S4

References 34-35

Materials and Methods

Participants and Procedure

To minimize the variance in norm compliance due to gender, only female undergraduate students at the University of Zurich participated in our study. The social experiment comprised 77 participants (mean age 22 +/- 0.4 [SEM] years, Range = 18 – 32 years) and the non-social experiment 64 participants (mean age 22 +/- 0.3 [SEM] years, Range = 18 – 32 years). For each experiment, participants were randomly assigned to one of three groups that differed only with respect to the type of transcranial direct current stimulation (tDCS) they received: anodal, sham, or cathodal; see the section "tDCS" and Table S1 further below for details. Participants in the three groups were well matched with respect to socioeconomic and personality variables; see the section "Analysis and Results" further below. Participants gave informed consent prior to the study. All experimental procedures were approved by the local ethics committee.

Testing was always performed in groups of 12 participants, except when some of the invited participants did not show up. However, in any case, the group of participants was randomly and evenly assigned to the three stimulation conditions. The experiment was conducted in the computerized group room of the Laboratory for Social and Neural Systems research (SNS-Lab). The group room comprises 14 identical computer workstations that are interconnected and shielded in view from one another, making it possible to conduct studies with anonymous, fully randomized social interactions (see Fig. S2). A multi-channel tDCS stimulator was used to simultaneously stimulate each of the 12 participants with anodal, sham or cathodal tDCS. Assignment to one of the three tDCS groups was performed in a double-blind fashion, with the participants and the experimenter who conducted the experiment not knowing which seats received active or sham stimulation. This group testing of participants thus controlled for unspecific effects, such as order, experimenter, and time of day effects that may potentially confound serial testing regimes.

Experimental Paradigm and Measures

Two weeks prior to the experiment, participants completed an online questionnaire containing several personality questionnaires measuring subjects' degree of Machiavellism (Mach IV scale), their risk-taking attitudes (DOSPERT scale) anxiety (STAI scale), and empathy (IRI scale). On

the day of testing, social norm compliance was measured using an experimental paradigm that closely follows the procedure used in a previous fMRI study (*10*). In this paradigm, participants repeatedly take the role of Player A and are randomly paired in every round with an anonymous player B. In every round both players receive an initial endowment of 25 money units (MUs). In addition, player A receives another 100 MUs that she can share with Player B as he likes. The sharing decision takes the form of a proposed transfer X from Player A to Player B and X can be any integer in steps of 10 between 0 and 100 (e.g., 0, 10, 20, etc., until 100). This decision is implemented by means of a visual analog scale and a computer mouse (see Fig. S1). In baseline roounds (see Fig. 1A in the main paper), player A proposes a transfer X which is always implemented as proposed. In punishment rounds (see Fig. 1B in the paper), by contrast, Player B has the option to use her initial endowment of 25 MUs to punish player A after she has observed the proposed transfer. In particular, for every MU that B invests into punishment player A's earnings are reduced by 5 MUs. This means, for example, that if player A transfers nothing to B such that after the transfer decision A has 125 MUs and B has 25 MUs, B can reduce A's earning to zero by investing the whole initial endowment into the punishment of A. During the experiment, each participant faced control trials and punishment trials in a random order, with the prevailing trial type indicated at the beginning of each round. In total player A completed 12 rounds in the baseline and 12 rounds in the punishment condition; in each round player A was matched with a randomly selected anonymous interaction partner.

The behavioral experiment described above took place in two separate contexts involving different groups of participants (see section "Participants and Procedures" above) – the social context and the non-social context. In the social context, Player A faced a different human interaction partner in every round. In punishment rounds of the social context, the human partner had the opportunity to punish Player A whenever she saw fit, for example, for unfairly low transfer levels. In contrast, in the non-social context, Player A was confronted with a pre-programmed computer. In punishment rounds of the non-social context the computer "punished" low transfer levels with exactly the same probability and magnitudes with which human partners punished low transfers in the social context.

At the beginning of each round the players were informed whether the upcoming trial belonged to the baseline or the punishment condition; this means that Player A always knew whether she faced a punishment threat or not. In baseline rounds, transfers therefore indicate

Player A's level of voluntary compliance with the fairness norm of an equal split. By contrast, Player A's sanction-induced norm compliance can be measured by the difference in transfer levels between punishment and baseline rounds, as this index quantifies how much the punishment threat makes Player A deviate from her level of voluntary norm compliance.

Given that the aim of this experiment was to test how tDCS affects Player A's norm compliance, players B were not physically present during the stimulation sessions but gave their responses in a pilot session recorded beforehand. However, all Players B agreed that their responses could be reused in other sessions (see also (*10*)). In the social context, each player A faced the decisions of a randomly selected player B and thus interacted with a real human opponent. All decisions were fully incentive compatible, as the MUs gained by the participants were transformed to Swiss Francs after the experiment according to a predefined conversion rate (1 MU = 0.015 CHF). These earnings were paid out on top of the base pay of 25 CHF (average pay = 88 CHF, max pay = 113 CHF, min pay = 52 CHF).

After participants had finished the behavioral paradigm (which lasted on average 11 minutes and 45 seconds), we measured several beliefs that the participants held about the paradigm while the tDCS stimulation was still ongoing. This was done to control for any possible effects tDCS may have on the representation of knowledge about the task or the opponent reactions. For these measures, participants reported in standardized questionnaires their beliefs about a) how fair Player A considers a transfer of 0, 20, 40 or 60 per cent of the endowment to be, b) how angry Player B would be when receiving a transfer of 0, 20, 40 or 60 percent and c) how strongly Player B would punish a transfer of 0, 20, 40 or 60 percent of the endowment. Responses to questionnaire a) and b) were given on a four-point scale ranging from "not at all" to "very", whereas responses to questionnaire c) were given in terms of expected punishment in MUs. This latter measure corresponded to the deduction resulting from Player B's response, e.g., 5*Y in Fig. 2B in the main paper.

Transcranial Direct Current Stimulation (tDCS)

During the experiment, we applied tDCS over the participant's rLPFC using a commercially available multi-channel stimulator that allows simultaneous stimulation of up to 16 participants with individually tailored stimulation protocols (see Fig. S2). tDCS modulates regional neural excitability by means of weak currents that increase or decrease the resting membrane potential,

depending on the position and polarity (anodal or cathodal) of the electrode. Thus, tDCS leads to an increase or decrease of the neural excitability in the brain tissue under the electrode (*17, 34*). In the present study, we applied anodal, cathodal, or sham tDCS over the right rLPFC region found activated in (*10*). The stimulation point was defined using the MNI coordinates reported by (*10*) as the group activation peak for the rLPFC region (x=52, y=28, z=14) that showed both heightened BOLD activity for punishment rounds minus baseline rounds as well as a correlation of individuals' BOLD activity with their transfer difference between punishment and baseline rounds. This standard coordinate was transformed to the individual head-space of each participant using T1-weighted MR scans of participant's neuroanatomy (T1-weighted 3D turbo field echo, 320 sagittal slices, matrix size: 240 x 240, voxel size = 1*1*0.6 mm, 8-channel head coil). The scalp coordinate overlying this brain area was employed as the center point for the target electrode and was determined for each participant prior to the experiment using Brainsight 2.0 frameless stereotaxy.

tDCS was applied using a set of standard 5x7 cm electrodes fixed by rubber straps. These standard electrodes were chosen over custom, more focal electrodes as we wanted to ensure that the large electrode would cover all neural rLPFC regions that are maximally active for each of our participants (minor variations in the precise spatial location of this area are averaged out during fMRI group analyses and are therefore likely to spread around the group peak used to define the stimulation site). The reference electrode (cathode for anodal tDCS and anode for cathodal tDCS) was positioned over the vertex, defined in the MR images as the scalp position overlying the confluence of each individual participant's right and left central sulcus. This reference electrode position thus circumvented influences on other cortical areas potentially relevant for the top-down control of behavior (e.g., other prefrontal regions). Importantly, the fMRI analyses reported in (*10*) did not reveal any activation in the vicinity of this reference electrode (e.g., in parietal cortex or posterior midline structures). We could therefore be confident that the effects of tDCS on norm compliance would not be mediated by neuromodulatory influences on task-related neural activity under the reference electrode. The only difference between the anodal and the cathodal group was therefore whether the anodal or the cathodal electrode was positioned over the rLPFC.

In line with established procedures, we stimulated with 1 mA current strength for the active anodal and cathodal groups. We accounted for possible delays in the onset of stable tDCS effects

(see (*17, 35*)) explicitly in our statistical analyses (see section "Analysis and Results" below). At the beginning of the stimulation the currents were slowly ramped up for 10 seconds to minimize tingling sensations caused by abrupt onsets of the tDCS. Likewise, when we finished brain stimulation the currents were slowly ramped down for 10 seconds. In the sham placebo group, the tDCS was turned off after 30 seconds. This latter condition feels identical to the active anodal and cathodal condition, but does not induce effects on neural excitability that outlast the stimulation period. The effects of the stimulation were indeed perceptually indistinguishable to the participants, as ascertained by a questionnaire conducted after the experiment in which the participants indicated how much they perceived the stimulation to affect their behavior (ranging from 1/"not at all" to 4/"extremely"). Participants in all three groups gave similar and statistically indistinguishable ratings (mean anodal: 1.42; mean sham: 1.30, mean cathodal: 1.55; $F(2, 62) = 0.24$, $p = 0.79$, ANOVA). Moreover, the different tDCS manipulations did not differentially affect the participants' general emotional state, as measured by three subscales of a standardized questionnaire (MDBF) indexing mood, alertness, and calmness. These scales were measured at the beginning and end of the experiment, while participants were still being stimulated with tDCS. Neither were there any differences with respect to variables before the test (ANOVA, all $F(2,60) < 0.34$, all $p > 0.71$) nor did their changes from before to after the test differ between the tDCS groups (all $F(2,60) < 1.64$, all $p > 0.2$). Taken together, these control analyses therefore show that unspecific non-neural effects of tDCS on beliefs about stimulation or general emotional state cannot explain changes in norm compliance due to the brain stimulation.

Analysis and Results

The randomization worked well with respect to balancing the groups for socioeconomic and personality variables. We computed ANOVAs with the factors experiment (social vs non-social) and tDCS (anodal, cathodal, and sham) to compare the different groups across the various measures acquired during the initial online questionnaire. None of these analyses revealed any significant main effects or interactions (see Table S1), showing that the groups were well matched with respect to variables other than tDCS that may have affected punishment-induced norm compliance.

Several participants had to be excluded from the analyses of their behavioral performance in the norm compliance paradigm as they evidently did not understand the task or chose to not participate in it. The criteria for this were: Failure to report answers within the response time of ten seconds on a majority of trials (11 participants) or stereotypical responses of not transferring any money on every single trial (7 participants). An additional participant in the cathodal group had to be excluded for moving the tDCS electrode during the experiment, resulting in abortion of stimulation while performing the task. This left 63 participants (19 anodal, 20 sham, and 24 cathodal) in the final analyses for the social experiment and 59 participants (21 anodal, 18 sham and 20 cathodal) for the analysis of the non-social experiment.

To assess the effects of anodal and cathodal brain stimulation on punishment-induced and voluntary norm compliance in the social experiment, we ran comprehensive generalized least-squares (GLS) regression analyses in STATA version 12. These analyses predicted for each individual *i* the observed choice $T_{i,t}$ in round *t* with the following equation:

$$T_{i,t} = \beta_0 + \beta_1 * \text{anodal} + \beta_2 * \text{cathodal} + \eta_i + \nu_t + \varepsilon_{i,t} \qquad \text{(eq. 1)}$$

For the analysis of voluntary norm compliance $T_{i,t}$ is given by the transfers in the baseline rounds. For the analyses of sanction-induced norm compliance $T_{i,t}$ is given by the difference between the transfers in the corresponding punishment and baseline rounds. Anodal and cathodal are dummy-coded variables that are set to 1 if individual *i* received anodal or cathodal stimulation, respectively, or to 0 in all other cases. Thus, the parameters β_1 and β_2 quantify the change in either voluntary or punishment-induced norm compliance due to anodal or cathodal

tDCS relative to the (omitted) sham group. The model furthermore contained a constant β_0, which measures the average transfer or transfer difference in the sham condition, a time-invariant error term η_i capturing unobserved characteristics of each participant *i*, a time-specific error term ν_t capturing the effect that a specific time period *t* may have on transfers and transfers differences, and a residual error term $\varepsilon_{i,t}$. As the two independent variables anodal and cathodal are between-subject variables that vary only across individuals, we employed a random-effects model with robust standard errors adjusted for clustering on the subject level.

As described in the main text, we find that anodal tDCS increases and cathodal tDCS decreases sanction-induced norm compliance and has opposite effects on voluntary norm compliance. These analyses focused on rounds 4-12 of the experiment, as it is well known from basic neurophysiological studies in humans (*17*) and mouse slice preparations (*35*) that the impact of tDCS on brain excitability becomes more robust and long-lasting after several minutes (this may possibly reflect delayed short-term neuroplastic processes occurring on top of immediate membrane potential changes; see (*17, 35*)). Thus, any corresponding behavioral effects of tDCS may also be expressed more profoundly several minutes after the onset of the tDCS. We accounted for this possible delay by focusing on rounds 4-12 which occurred after a minimum of 5 minutes after the onset of the tDCS and therefore lie fully in the temporal window where tDCS exerts lasting neurophysiological effects (see (*17, 35*)). However, we ensured that the precise cut-off points for including periods into the analyses did not affect the results, by conducting control analyses in which we also included data from earlier periods, or where the cutoff point for inclusion was moved to later periods. Table S2 and S3 show that very similar tDCS effects are obtained if we include previous periods or start the analysis at later periods. This shows that the changes in sanction-induced and voluntary norm compliance resulting from tDCS are temporally robust and do not depend on the particular time window (periods 4-12) we have chosen.

As the critical comparison between the three tDCS conditions comprised different groups of participants, one may wonder whether possible differences in the personality of the participants may have contributed to the group differences in sanction-induced norm compliance. Table S1 already shows that this scenario is very unlikely, as such personality variables did not differ between the three tDCS groups due to our strict randomization procedures. We

nevertheless directly controlled for these personality variables in our statistical analysis, by repeating the regression model given in eq.1, but now adding all personality variables given in Table S1 as additional regressors. This control analysis revealed very similar parameter estimates and significance levels for the effects of anodal (sanction-induced norm compliance: 9.9282., $p<0.0001$; voluntary norm compliance: -7.97, $p < 0.0001$) and cathodal (sanction-induced norm compliance: -8.128, $p<0.0001$; voluntary norm compliance: 3.028, $p < 0.024$) tDCS. This demonstrates that it was indeed the tDCS – and not any possible (non-significant) differences in personality characteristics – that led to different levels of sanction-induced and voluntary norm compliance in the different groups.

The results so far show that tDCS of rLPFC affected sanction-induced and voluntary norm compliance, but they do not yet show whether these effects relate specifically to the *social dimension* of the interaction. This question appears somewhat irrelevant for our findings on voluntary norm compliance, as fairness norms prescribing voluntary sharing of money only exist for social interactions with other humans. However, this question is relevant for understanding the tDCS effects on sanction-induced norm compliance, as transfer decisions in the punishment condition do not only require social/normative considerations. For instance, transfer choices in this condition require both the assessment of the risk of being punished as well as the evaluation of the trade-off between selfish gains from low tranfers and the associated loss from possible punishment; any tDCS effect on such generic decision processes may influence behavioral reactions to sanction threats independently of the social nature of the punishment. To examine whether the reported tDCS effects on sanction-induced norm compliance are indeed specific to the *social* context, we therefore repeated the experiment in a new sample of participants who participated in the same economic game as before, but now played against a computer that was pre-programmed to respond to the transfers in the same way as a human opponent. For this purpose, we determined the computer's response on every trial by a random draw from the actual distributions of punishment choices from the first experiment for a given transfer. Adding these data to the regression model specified in eq.1 yields the following full model:

$$T_{i,t} = \beta_0 + \beta_1 * \text{anodal} + \beta_2 * \text{cathodal} + \beta_3 * \text{non-social} + \beta_4 * \text{anodal} * \text{non-social} + \beta_5 * \text{cathodal} * \text{non-social} + \eta_i + \nu_t + \varepsilon_{i,t} \quad \text{(eq. 2)}$$

The variable "non-social" is dummy-coded, i.e., it is set to 1 if participant *i* played against a computer and 0 if she played against a human player. Thus, the crucial parameters β_4 and β_5 specify how the tDCS-effects on sanction-induced norm compliance change when participants are faced with a computer rather than a real person. All other variables are coded in the same way as in eq. 1. The regression results for model (2) are given in Table S3. Crucially, the significant interaction parameters β_4 and β_5 show that the effects of social sanction threats on norm compliance were indeed much stronger than the corresponding effects in the nonsocial experiment, with anodal tDCS leading to a stronger increase and cathodal tDCS to a stronger decrease in sanction-induced norm compliance. Again, these effects were robust across different time windows of the experiment and were unaffected when controlling for potential effects of personality variables in the statistical model. Taken together, these analyses demonstrate that the rLPFC indeed plays a specific role in integrating the social dimension of possible sanctions into behavioral control based on social norms.

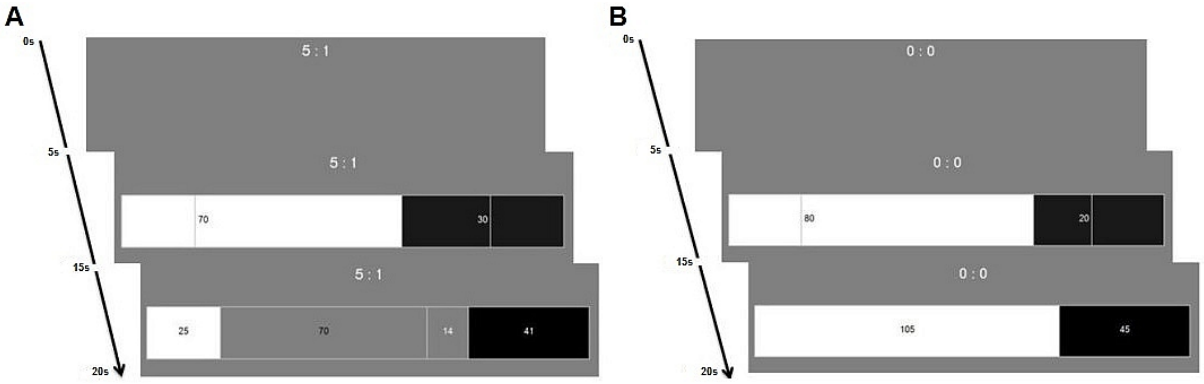

Figure S1. Behavioral paradigm. Schematics of the visual displays used for the experimental task in a punishment round (**A**) or baseline round (**B**). Participants were first informed by means of a visual cue (presented for 5 seconds) whether punishment was possible or not. This was indicated by the numbers 5:1 (as in panel A, specifying that each punishment point invested by B reduced Player A's monetary payoff by 5 points) or 0:0 (as in panel B, no points could be invested by B and deducted from A). This was followed by a bar stimulus (presented for 10 seconds) used by Player A to indicate the designated transfer, by moving a computer mouse to the corresponding position and confirming the selection by mouse click. The length of the white bar indicated the portion of the 100 MUs Player A wanted to keep for herself, whereas the length of the black bar indicated the transferred amount (these choices were also displayed numerically at the left and right end of the bar). For example, in the left figure above Player A kept 70 MUs for herself and transferred 30 to Player B. The 25 MUs given to Players A and B were also displayed as separate sections of these bars on either end that could not be altered by player A's choice. Following a break of 5 seconds in which player B's choice was determined, the final outcome was revealed. In the baseline trials, this corresponded to the proposed split, whereas in punishment rounds, the money deducted from both Player A and B as a consequence of Player B's punishment choice was displayed, by overlaying a grey bar over the white and black bars indicating Player A's and B's payoffs. The size of this bar corresponded to the amount of punishment B decided to impose on A following the 5:1 punishment ratio. The final outcome of the interaction was thus displayed in the re-sized white bar (for Player A) and black bar (for Player B). For example, in the left figure above, after Player A chose a 70:30 split of the 100 MUs, Player B invested 14 into punishment which led to a final payoff of 41 for B. The sanction imposed by B reduced A's payoff by 5*14 = 70 MUs, leaving A with a final payoff of 25MUs.

Figure S2. tDCS Setup.

All testing was conducted in sessions with 12 participants (except when not all invited participants showed up) who were randomly and evenly sorted into the three stimulation conditions (4 anodal, 4 cathodal, and 4 sham) in a double-blind design. Participants were seated in a group laboratory, each facing an identical computer workstation that was shielded from the other players' view. tDCS was employed via a multi-channel tDCS stimulator that can apply individualized electric current stimulation protocols to the brain of each volunteer (see text). This parallel testing regime has the advantage that many unspecific testing effects (e.g. time of day, experimenter, etc.) are identical for the different stimulation groups.

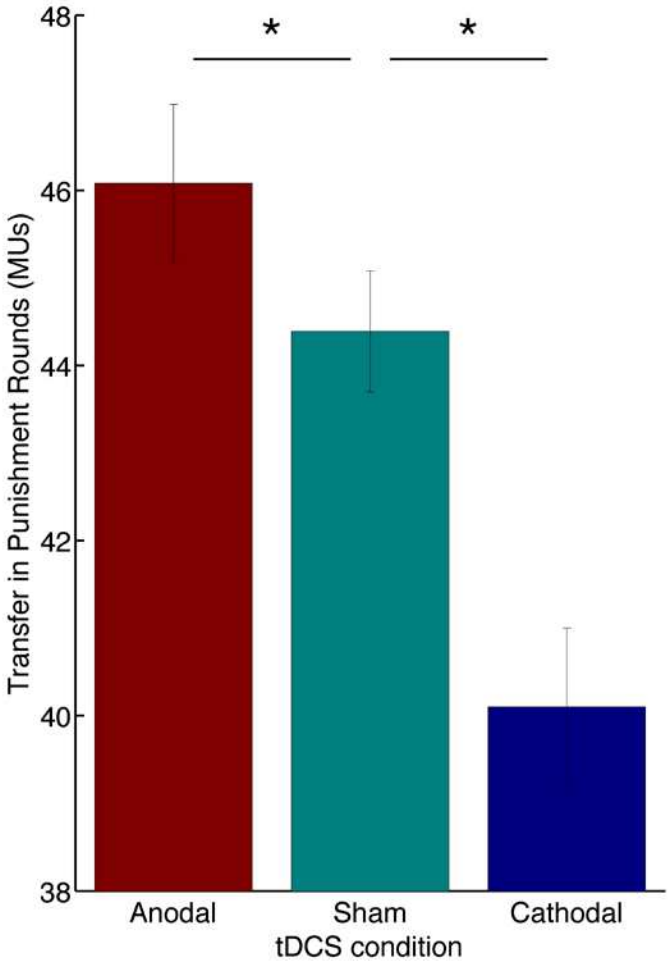

Figure S3. Anodal tDCS increases and cathodal tDCS decreases transfers in punishment rounds. The bars depict Player A's average transfers (+/- s.e.m.) across all punishment rounds. Note, however, that the effect of sanction threats on norm compliance can only be accurately quantified for each individual in relation to her level of voluntary norm compliance in the baseline rounds. The individual transfer difference between punishment and baseline rounds (Fig. 2A) is therefore used as index of sanction-induced norm compliance. All values determined with regression in eq. 1; * $p < 0.05$.

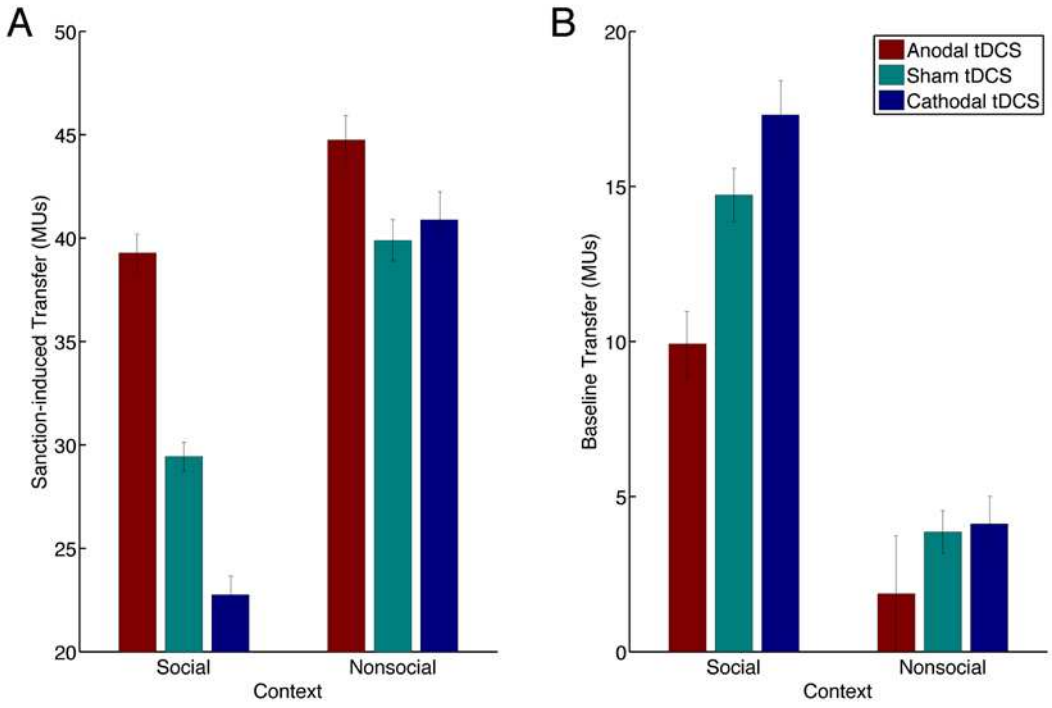

Figure S4. tDCS effects on sanction-induced and voluntary transfers are more pronounced in the social context than in the non-social context. (A) The bars depict Player A's average transfer difference (+/- s.e.m.) between punishment and baseline rounds. Positive values indicate higher transfers when sanction threats are present. In the non-social context, participants were clearly also sensitive to punishment threats (as indicated by the large positive values), but the effect of tDCS on sanction-induced transfers (difference between the three bars in different colors) was much weaker than in the social context; see Fig. 4 and main text for direct comparison and statistics. All values determined with regression in eq. 2. **(B)** The bars depict Player A's average transfer (+/- s.e.m.) in baseline rounds. In the non-social context, participants hardly transferred any MUs to the computer opponent, whereas in the social context, they voluntarily transferred MUs to the anonymous human opponent. Again, the effect of tDCS on voluntary transfers (difference between the three bars in different colors) was more pronounced in the social context; see Fig. 4 and main text for direct comparison and statistics. All values determined with regression in eq. 2.

Tables

		Machiavelli		Dospert		Davis		STAI	
Experiment	Factors	F Value	P Value	F Value	P Value	F Value	P Value	F Value	P Value
Social	tDCS	1.36	0.27	1.17	0.32	0.1	0.91	0.28	0.75
Non-social	tDCS	0.89	0.42	0.78	0.46	0.68	0.51	0.82	0.45
Combined	tDCS	1.37	0.26	1.26	0.29	0.09	0.91	0.24	0.78
	Experiment	2.44	0.12	2.03	0.16	1.17	0.28	0.82	0.37
	tDCS*Experiment	2.16	0.12	1.68	0.2	0.53	0.59	0.98	0.38

Table S1: Summary of statistics testing for possible stimulation group differences in variables that may have affected sanction-induced norm compliance. We compared scores on several personality scales measuring Machiavellian thinking, risk attitudes (Dospert scale), empathy (Davis scale), and anxiety (STAI) for participants in the social experiment (with human opponents) and the non-social experiment (with computer opponents). For this purpose, we conducted ANOVAs with the independent variables tDCS (3 levels: anodal, sham, and cathodal) and experiment/context (2 levels: social and non-social experiment/context). This revealed that the groups did not differ in any of these variables, making it unlikely that differences in personality variables could account for the effects of tDCS on sanction-induced and voluntary norm compliance.

	Periods 1-12		Periods 2-12		Periods 3-12		Periods 4-12		Periods 5-12		Periods 6-12		Periods 7-12	
Regressor	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value
Andoal	7.204	0.0001	8.126	0.0001	9.24	0.0001	9.854	0.0001	9.539	0.0001	9.431	0.0001	9.873	0.0001
Cathodal	-6.253	0.0001	-6.471	0.0001	-6.321	0.0001	-6.691	0.0001	-6.67	0.0001	-7.105	0.0001	-6.602	0.0001
Constant	28.8	0	29.28	0	29.28	0	29.44	0	29.93	0	30.4	0	30.37	0

Table S2: Sanction-induced norm compliance: Summary statistics for the GLS regression given in eq. 1. The analysis estimates the impact of anodal and cathodal stimulation on transfer difference between punishment and baseline rounds across different time windows. To ensure that we capture the time window during which tDCS exerts lasting neurophysiological effects we concentrate our analysis in the paper on periods 4-12. Here we show the regression coefficients for our tDCS effects (anodal, cathodal) for larger and smaller time windows around our preferred window (4-12). The effects remain stable and significant across all the above time windows, indicating temporally robust effects of tDCS on behavior.

	Periods 1-12		Periods 2-12		Periods 3-12		Periods 4-12		Periods 5-12		Periods 6-12		Periods 7-12	
Regressor	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value
Andoal	-3.303	0.001	-4.072	0.001	-4.732	0.001	-4.805	0.001	-4.739	0.001	-4.333	0.001	-4.35	0.001
Cathodal	1.549	0.0423	2.030	0.0977	2.150	0.053	2.870	0.0093	2.635	0.0257	3.107	0.0118	3.167	0.02
Constant	10.42	0.001	15.55	0.001	15.10	0.001	14.72	0.001	14.5	0.001	14.21	0.001	4.048	0.001

Table S3: Voluntary norm compliance: Summary statistics for the GLS regression given in eq. 1. The analysis estimates the impact of anodal and cathodal stimulation on voluntary transfers in baseline rounds across different time windows. We again report this analysis in the paper for periods 4-12 to account for possible delays in the onset of stable neurophysiological effects due to tDCS. Here we show the regression coefficients for our treatment effects (anodal, cathodal) for larger and smaller time windows around our preferred window (4-12). The effects remain stable and significant across all the above time windows, indicating temporally robust effects of tDCS on behavior.

	Periods 1-12		Periods 2-12		Periods 3-12		Periods 4-12		Periods 5-12		Periods 6-12		Periods 7-12	
Regressor	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value	Coeff.	P Value
Andoal	7.209	0.0001	8.130	0.0001	9.243	0.0001	9.859	0.0001	9.543	0.0001	9.435	0.0001	9.873	0.0001
Cathodal	-6.255	0.0001	-6.471	0.0001	-6.321	0.0001	-6.691	0.0001	-6.67	0.0001	-7.105	0.0001	-6.602	0.0001
Non-Social	9.981	0.0001	10.31	0.0001	10.2	0.0001	10.45	0.0001	9.712	0.0001	9.535	0.0001	9.709	0.0001
Anodal*Non-Social	-1.628	0.424	-2.603	0.1650	-4.011	0.0490	-5.009	0.0091	-4.452	0.038	-5.013	0.0397	-5.670	0.0246
Cathodal*Non-Social	8.061	0.0001	7.530	0.0002	7.445	0.0005	7.686	0.0005	8.277	0.0002	8.601	0.0003	7.356	0.0036

Table S4. Summary statistics for the GLS regression in eq. 2 that estimates whether the impact of anodal and cathodal stimulation on sanction-induced norm compliance is stronger in the social context than in the non-social context. The significant interaction terms confirm that tDCS effects on sanction-induced norm compliance were indeed stronger during social interactions with a human opponent. Again, we performed these analyses for various time windows to account for possible delays in neurophysiological tDCS effects. Except for the inclusion of the first two periods, the interaction effect is always significant for anodal stimulation, suggesting that the effect builds up over time; for cathodal stimulation is it even significant if we include the first two periods in the analyses.